Text Mining Summary Statements for Evidence of Innovation

Daniel Russ[1], Calvin Johnson[1], and Luci Roberts[2]
[1]Division of Computational Bioscience, Center for Information Technology
[2]Division of Planning and Evaluation, Office of Extramural Research

Abstract

Questions about the NIH research portfolio can be difficult to address in cases where the answers are not stored as structured fields in a database. Inferences can be made to address such questions using text analysis, but robust text analysis methods for dynamically testing hypotheses about the NIH research data are lacking. *Innovation* is one of five review criteria addressed in the initial peer review of NIH research grant applications. Reviewers' narrative critiques of innovation are contained in summary statements. We present a study on the use of text mining to identify applications that peer reviewers assessed as innovative. Specifically, we sought to address the following questions: Is there a relationship between the scientific review group's assessment of overall impact and the individual reviewers' narrative descriptions of innovation? Are there differences in innovation related to the career stage of the investigator? Are there individual differences among investigators in reviewers' assessments of innovation? To develop a training set, we asked NIH Scientific Review Officials to select text from summary statements that indicated innovation (or lack of innovation) on a 5-point scale. Using the annotated text, we built a lexicon of words and phrases that describe innovation, and developed a classifier that can select innovative documents. The lexicon that emerged was quite limited, consisting of only a short list of terms. However, these terms were found to have relatively strong utility in predicting innovation, based upon its relationship with criterion scores for innovation. We also identified a significant relationship between reviewers' positive sentiments about innovation and favorable priority (overall impact) scores. This relationship held for both old and new ("enhanced") critique formats. We identified no meaningful differences between new and established investigators in reviewers' sentiments about innovation, but we found that New Investigators whose applications were described by reviewers as innovative were significantly more likely to submit subsequent applications that were also described by reviewers as innovative. Thus, New Investigators who are identified as innovative are more likely to be innovative in the future.

Methodology

Annotation

Twelve NIH Scientific Review Officials annotated 115 NIH summary statements selected because the application was considered strongly innovative or as a negative training document. The summary statements were broken into sections, only the résumé and critiques were analyzed. The résumé and critiques are collectively referred to as documents for this experiment. Annotation software was provided to the Scientific Review Officials, who collected document level and phrase level annotation explaining why the Scientific Review Official labelled the section as innovative. The document and the selected phrases are scored on a 5 level scale (very negative innovation, negative innovation, neutral, positive innovation, very positive innovation). Figure 1 shows a screen shot of the annotator.

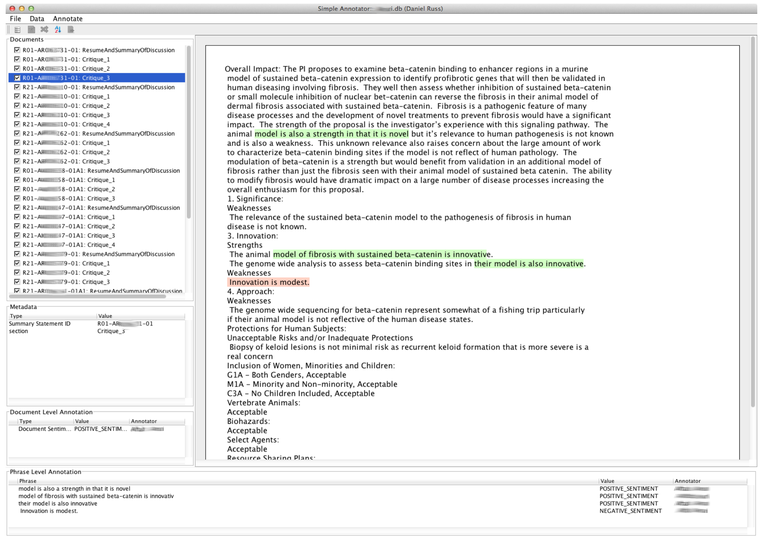

Figure 1: Screen shot of our annotator software. Scientific Review Officials can select phrases that they believe reflects innovation (or lack of innovation). Both phrase and document level annotation is collected. The annotation is saved in a local database that can be merged with annotation from other annotators for portfolio analysis.

Text Mining

The text annotations were grouped by hand to reflect the similar syntactic structure of the annotation. These similarities were encoded using regular expression to easily identify instances in text. The regular expressions formed a set of rules that were used as classification features for an innovation classifier. Figure 2 shows two rules used in the text mining. As an example, the phrase "*The central hypothesis is extremely innovative*" is very similar to other phrases that use "*is very innovative*". The rule looks for the "to be" verb, potentially followed by one or more adverbs, followed by the word innovative. However, innovative has synonyms that needed to be considered. Using the dictionary, we developed a list of synonyms that are shown on the Figure 2. The second rule handles phrases like "*innovative experimental approach*". Rules can also identify comments that are not innovative, such as "*described 12 years ago*". Adverbs that negated or limited the innovation, such as "*moderately innovative*", "*nothing innovative*", or "*not particularly innovative*" were marked as negated rules.

An overview of the classification is shown in Figure 3. Given a set of rules, each rule that matches a phrase is recorded in a matrix, referred to as a Document-Rule matrix. The matrix is used to train a Maximum Entropy classifier. The classification results are compared to the document level annotation provided by the Scientific Review Official. The performance of the classifier was examined, changes were made to the rules and the classifier re-trained. After the training process, the classifier was tested on a second corpus of summary statements. Two hundred documents were selected out of the un-amended, type 1 applications submitted to the May 2012 NIGMS, NCI and NIAMS Councils. The trained classifier was compared against the document level annotation.

Comparison between Established Investigators and New Investigators

Three corpora were created for this analysis. The 2004 New Investigator corpus contained ~~the~~ critiques from applications awarded in 2004 - 2009 from New Investigators whose first R01 awards were made in FY 2004. These data were examined to explore whether there were differences in reviewers' ratings of innovation as a function of reviewers' assigned role (resume, primary, secondary, tertiary reviewer and beyond), application activity code, or

Rule: innovative (adj)

(verb)	(adverb)*	***innovative***	(adj|adv)*	(noun|verb)
is	very	innovative		
		innovative	technical	approach
		novel	regulatory	networks

innovative = innovative, imaginative, paradigm-shifting, groundbreaking, pioneering, straightforward, controversial, incremental, fresh, state-of-the-art, novel, creative, seminal, revolutionary, cutting-edge

Figure 2: Example of rules for identifying innovative sentences. Sentences containing any adjective in the category ***innovative*** can be found in one of two contexts. First, following the verb "to be" potentially having an adverb. Second, preceding a noun potentially having an adjective. The optional adverbs are check for potential negation/down-weighting (e.g. weakly, somewhat)

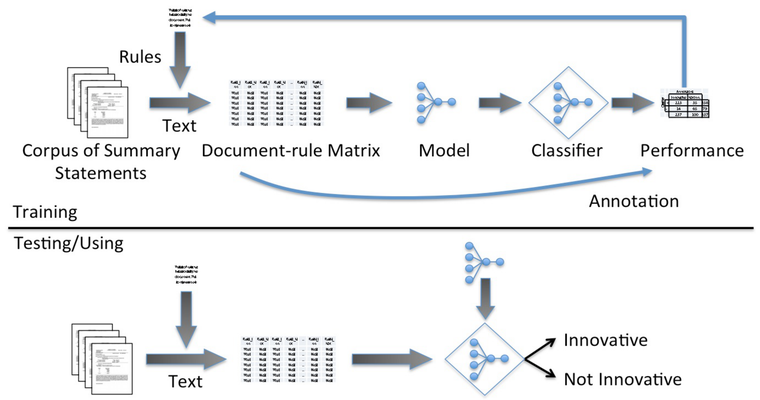

Figure 3: Overview of the text mining paradigm. In the top of the figure, text is extracted from a training corpus. An initial set of rules is applied to train a maximum entropy classify. The performance of the classifier is examined to decide if addition rules need to be added to the classifier. When the performance is adequate, text is extracted from an additional testing corpus. The model calculated the probability that the text is innovative or not innovative. The document is classified with the most probable label.

fiscal year. No significant differences were identified in innovation as a function of activity code or fiscal year, but the resume and critiques of primary and secondary reviewers were found to be significantly more likely to identify innovation in the applications in comparison to tertiary reviewers, and secondary critiques were less innovative than those of primary reviewers, but the difference was not significant. For this reason, subsequent analyses focused on innovation levels assessed in the résumé, primary and secondary critiques. The 2004 New Investigator corpus was narrowed to only those investigators whose subsequent competing continuation applications were awarded and the level of innovation assessed in the initial application in comparison to the competing continuation from the same investigator. The 2009 established investigator corpus contained 5929 R01 awards to these investigator until 2009. The 2009 New Investigator corpus consists of 1572 R01 awards to New Investigators. Logistic regression analysis was used to assess the contribution of these three factors and predict the level of innovation.

Results

Classifier

The classifier performed very well against the test dataset when the neutrals were removed. The recall was 91% and the precision was 75%. However, with the neutrals, the precision fell to 55%, the recall was unaffected.

Comparison between Innovation Score and System classification

Figure 4 shows the predicted probability that an application is innovative as a function of the innovation criterion score assigned by the reviewers. The predicted value is given by the logistic regression analysis, the observed values were calculated from text mining. Some critiques contained assessments of innovation that were rated by the classifier inconsistently with the innovation criterion score assigned by the reviewer. Overall, however, the innovation rating assigned by the text-mining classifier was found to significantly predict the level of innovation reflected by the innovation criterion score.

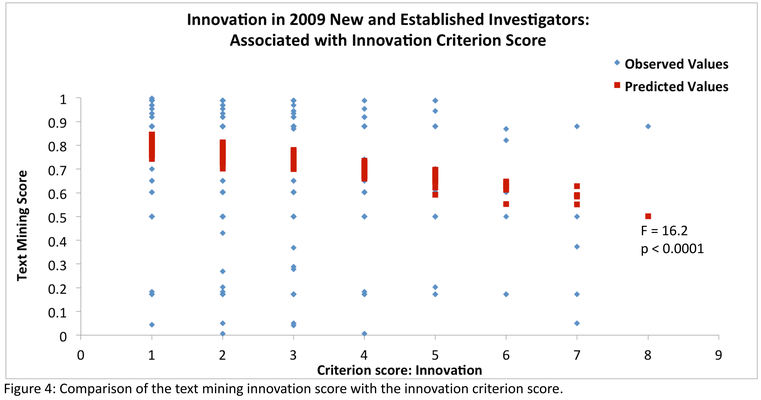

Figure 4: Comparison of the text mining innovation score with the innovation criterion score.

Note: Since the text mining score measure is limited to a range of 0 to 1, an arcsine square root transformation was applied to the data. The analysis of the transformed data did not yield results that were different from those of the analysis of the untransformed data. The graphical presentation of the results depicts the analysis of the untransformed data.

A weak but significant relationship was found between reviewers' assessments of innovation assigned by the classifier and the priority scores in FY 2004 (Figure 5). Similarly, the relationship between innovation level assigned by the classifier and the Overall Impact score assigned by reviewers to the applications awarded in FY 2009 was also significant (Figure 6).

Finally, New Investigators whose applications were rated to be innovative in 2004 were significantly more likely to submit subsequent applications that were also considered innovative, as shown in Figure 7.

Discussion

As an example on the utility of text mining to perform portfolio analysis, we attempted to build a classifier that could identify innovation from critiques and résumés in NIH summary statements. We were able to identify phrases that implied an application was considered innovative. This information was useful to classify applications as innovative or not. The innovation classification was useful to classify almost 8800 critiques, applying a uniform set of classification rules to uniformly evaluate the critiques. If performed manually, such an analysis would have been a large burden on program staff.

The system was able to identify innovative applications with good recall and precision. Summary statements without comments on innovation are difficult for the system to assess, and many system error occurs because of this situation.

However, we were able to show that as the innovation criterion score decreases the text mining innovation score increase. As the priority (2004) or impact score (2009) improves so does the innovation score. Finally, we were able to show that innovative investigators remain innovative later in their career.

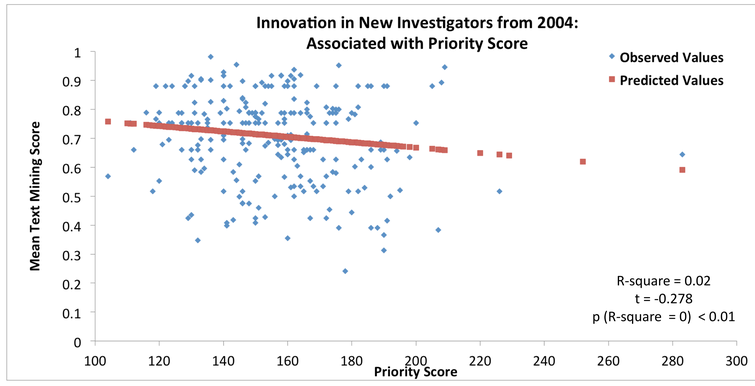

Figure 5: Comparison of the mean text mining innovation score for an application with the priority score for 2004 New Investigators.

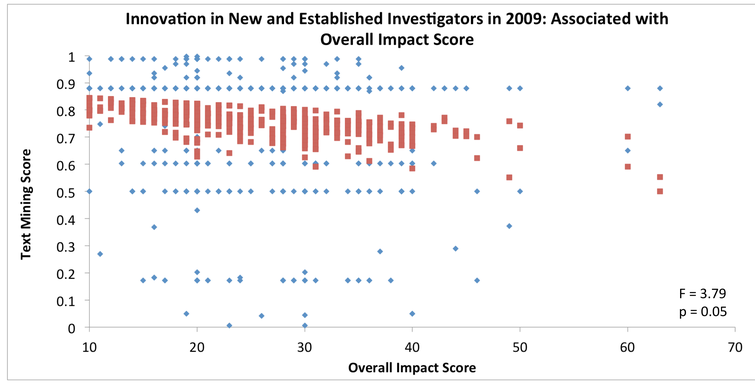

Figure 6: Comparison of the text mining innovation score with the overall impact score for 2009 New Investigators.

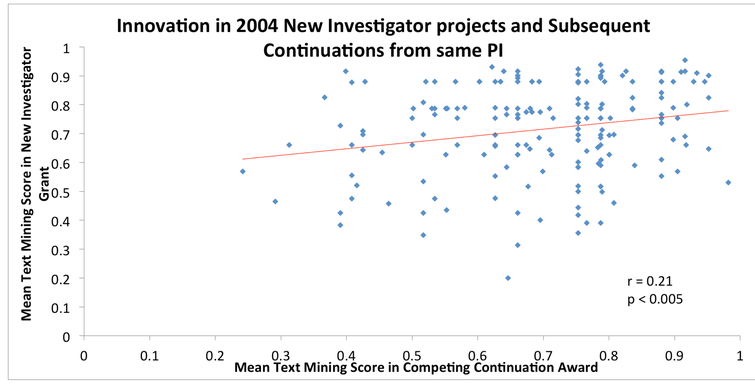

Figure 7: Comparison of the text mining innovation score for 2004 New Investigators and subsequent application from the same PI.